Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Научно-образовательный центр когнитивного моделирования

Направление подготовки / специальность: 03.04.01 Прикладные математика и физика
Направленность (профиль) подготовки: Методы и технологии искусственного интеллекта

РАЗРАБОТКА НЕЙРОСЕТЕВОГО МЕТОДА РАСПОЗНАВАНИЯ РЕЧИ В УСЛОВИЯХ СИЛЬНОЙ ЗАШУМЛЕННОСТИ

(магистерская диссертация)

Студент:
Петряшин Илья Евгеньевич



(подпись студента)

Научный руководитель:
Юдин Дмитрий Александрович,
канд. техн. наук

_____
(подпись научного руководителя)

Консультант *(при наличии)*:
_____
_____
(подпись консультанта)

Москва 2024

Содержание

1 Введение **4**

2 Обзорно-постановочный раздел работы **5**

2.1 Основные понятия и определения 6

2.2 Методы детекции голосовой активности 6

2.3 Методы распознавания речи 7

2.4 Методы денойзинга речи 12

3 Постановка задачи **14**

3.1 Детекция голосовой активности в режиме реального времени 14

3.2 Распознавание речи в условиях сильной зашумленности 15

4 Метод **16**

4.1 Детекция голосовой активности в режиме реального времени 16

4.2 Распознавание речи в условиях сильной зашумленности 19

4.2.1 Денойзинг входной последовательности 19

4.2.2 Коррекция выходной транскрибации 21

5 Вычислительные эксперименты **22**

5.1 Детекция голосовой активности в режиме реального времени 22

5.1.1 Входные данные 22

5.1.2 Анализ результатов 22

5.1.3 Выводы 24

5.2 Распознавание речи в условиях сильной зашумленности 26

5.2.1 Входные данные 26

5.2.2 Дообучение моделей распознавания 27

5.2.3 Использование недообученных денойзеров 28

5.2.4 Дообучение денойзеров 29

5.2.5 Использование непредобученных денойзеров 30

5.2.6 Дообучение денойзеров с использованием дополнительной функции потерь 31

5.2.7 Коррекция транскрибации 33

5.2.8 Генерализация денойзинга на других датасетах 34

5.2.9 Генерализация денойзинга на других моделях 36

5.2.10 Генерализация денойзинга на разные SNR 37
5.2.11 Выводы 38

6 Заключение **40**

Аннотация

Распознавание речи является важной задачей и имеет множество приложений в таких областях как телефония, робототехника и другие. Одной из главных проблем для распознавания речи является наличие фонового шума, который приводит к ухудшению качества распознавания, что особенно критично в условиях сильной зашумленности, в которых существующие модели распознавания показывают плохое качество.

В данной работе рассматривается задача распознавания речи в условиях сильной зашумленности для случая, когда необходимо распознавать речь в онлайн–режиме. Данная задача разбивается на 2 задачи: детекцию голосовой активности в режиме реального времени и распознавание речи в условиях сильной зашумленности, при этом большее внимание уделяется именно распознаванию.

Для решения задачи детекции голосовой активности предложен алгоритм RealVADR, позволяющий использовать существующие нейросетевые модели детекции голосовой активности в реальном времени. Исследовано, как параметры разработанного алгоритма влияют на качество распознавания и показано, что данный алгоритм позволяет получать качество распознавания, сравнимое с офлайн распознаванием.

Для решения задачи распознавания речи в условиях сильной зашумленности разработан оригинальный подход DASR, суть которого состоит в добавлении дополнительной нейросетевой модели-денойзера, которая обучается на выходе модели распознавания и изменяет только входной аудиосигнал. Исследовано, как предложенный подход влияет на качество распознавания и показано, что он позволяет достигать максимально достижимого качества. Кроме того, данный подход хорошо обобщается на множество речевых доменов, а также на другие модели распознавания. Также предложено использование моделей коррекции транскрибации, которые в сумме с денойзингом позволяют улучшать качество распознавания почти в два раза.

1 Введение

Распознавание речи в условиях сильной зашумленности представляет собой важную задачу в современном мире и имеет множество приложений. В различных областях, таких как телефония, автомобильная промышленность, домашние устройства и робототехника, точность и эффективность распознавания речи являются ключевыми факторами успешного внедрения технологий. Однако, наличие фонового шума часто приводит к значительному снижению качества распознавания, что делает эту проблему особенно актуальной.

Одной из основных технических проблем, с которой сталкиваются системы распознавания речи, является наличие фонового шума. Фоновый шум может быть вызван различными источниками, такими как шум ветра, шум различных приборов, разговоры других людей или звуки окружающей среды. Этот шум затрудняет правильное распознавание и интерпретацию речевых сигналов, что существенно снижает эффективность системы в реальных условиях эксплуатации.

Основные задачи в области распознавания речи включают в себя не только сам процесс преобразования аудиосигнала в текст, но и предварительную детекцию голосовой активности. Детекция голосовой активности (Voice Activity Detection, VAD) направлена на определение временных интервалов, в которых присутствует речь, и отсечение фонового шума и тишины. Эта задача решается с использованием различных методов машинного обучения, включая нейросетевые подходы, которые обеспечивают высокую скорость и точность работы в реальном времени.

Для детекции голосовой активности часто применяются методы машинного обучения, включая нейросетевые подходы. Эти алгоритмы обучаются на больших объемах аудиоданных, чтобы точно определять начало и конец речевых фрагментов в реальном времени. Использование таких алгоритмов позволяет значительно повысить точность и скорость работы системы распознавания речи.

Однако, даже после успешной детекции голосовой активности, качество распознавания речи может быть существенно снижено из-за фонового шума. Для улучшения качества распознавания в условиях сильной зашумленности применяются специализированные методы, такие как дообучение моделей распознавания на зашумленных датасетах или использование хорошо генерализованных моделей распознавания.

Перспективными для улучшения качества распознавания выглядят методы, использующие нейросетевые модели-денойзеры, которые направлены на устранение

шума и повышение различимости речевых сигналов.

Кроме того, в данный момент активно разрабатываются модели коррекции транскрибации, которые позволяют улучшать качество распознавания путем анализа и исправления ошибок, допущенных на предыдущих этапах обработки речевого сигнала.

В данной работе предложен комплексный метод к распознаванию речи, состоящий из двух частей: детекция голосовой активности в реальном времени и распознавание речи в условиях сильной зашумленности.

Для детекции голосовой активности в реальном времени разработан алгоритм RealVADR [39], использующий в своей основе существующие нейросетевые модели детекции голосовой активности и показано, что его применение в реальном времени позволяет получать схожеее с офлайн-распознаванием качество.

Для задачи распознавания речи в условиях сильной зашумленности разработан оригинальный подход DASR, основанный на использовании нейросетевых моделей денойзинга, которые изменяют входной аудиосигнал таким образом, чтобы повышалась его различимость моделями распознавания речи. В экспериментах показано, что предложенный подход позволяет достигать максимально достижимого качества распознавания, и при этом достаточно только изменения аудиосигнала без необходимости изменения модели распознавания. Также проведены эксперименты по оценки генерализации данного метода и показано, что он хорошо работает не только на других датасетах, но и позволяет улучшать качество даже тех моделей распознавания, для которых он не был изначально приспособлен, что делает его универсальным инструментом в условиях сильной зашумленности.

Также в работе рассмотрено применение различных моделей коррекции транскрибации, которые совместно с денойзингом способствуют значительному улучшению качества распознавания, повышая его почти в два раза.

2 Обзорно-постановочный раздел работы

В данной работе задача распознавания речи в условиях сильной зашумленности разделена на две подзадачи – задачу выделения из входящего аудиопотока интервалов, содержащих голосовую активность, и задачу улучшения качества распознавания в условиях сильной зашумленности.

2.1 Основные понятия и определения

VAD (Voice Activity Detection) – детекция голосовой активности в аудиопотоке

ASR (Automatic Speech Recognition) – распознавание речи

WER (Word Error Rate) – ошибка распознавания слов в речи

CER (Character Error Rate) – ошибка распознавания символов в речи

SNR (Signal to Noise Ratio) – отношение уровня шума к уровню сигнала

Транскрибация – речь, преобразованная в текст с помощью модели распознавания

Денойзер (модель денойзинга) – нейросетевая модель, предназначенная для удаления из сигнала шума или другого преобразования, улучшающего качество аудиосигнала

Генерализация – способность нейросетевой модели демонстрировать высокое качество работы не только на тех данных, на которых она обучалась, но и на новых, ранее не виденных данных, сохраняя при этом точность и надежность своих предсказаний

2.2 Методы детекции голосовой активности

Задача детекции голосовой активности представляет собой задачу выделения из входящего аудиопотока интервалов, содержащих голосовую активность. Трудность решения данной задачи заключается в том, что записанные аудиосигналы включают разнообразные виды контента помимо речи: окружающий шум, интервалы тишины и другой невербальный контент. Часто важно, чтобы на длинных интервалах речи распознавание проводилось только на сегментах, содержащих речевые данные, чтобы модели распознавания обучались и тестировались только на релевантной информации. Кроме того, некоторые речевые характеристики (такие как время начала и окончания высказывания или паузы между сегментами, относящимися к одному высказыванию) очень важны в различных приложениях, таких как обнаружение обманной речи [14]. Таким образом, первой задачей, которую необходимо решать в задачах обработки речи, является обнаружение голосовой активности (VAD), то есть точное определение временных интервалов речи в аудиозаписях.

Существующие техники в решении данной задачи включают использование алгоритмически извлеченных признаков речи, таких как энергия, частота пересечения нуля (ZCR), гармоническое распределение и отношение сигнал/шум, коэффициенты

перцептивного линейного предсказания (PLP) или мел-кепстральные коэффициенты (MFCC), а также различные спектральные признаки и их временные производные, в сочетании с моделями машинного обучения, такими как скрытые марковские модели (HMM), гауссовские смеси (GMM) или гибридные модели GMM-HMM [19, 6], фильтры Калмана [4], или, нейронные сети, включая многослойные перцептроны (MLP) [18] и рекуррентные нейронные сети (RNN), в том числе с использованием ячеек долговременной краткосрочной памяти (LSTM) [15].

Наиболее современные подходы используют сверточные нейронные сети (CNN) вместе с автоматически извлеченными признаками, как непосредственно из временных выборок, так и с использованием частотного представления сигнала, такого как дискретное преобразование Фурье (DFT) или спектрограммы [5]. Кроме того, наиболее хорошие результаты показывают совместно обученные модели CNN-MLP или CNN-RNN [13, 8].

Для практического использования доступны открытые реализации, такие как Silero [1] и pyannote [2]. Silero предоставляет набор моделей для VAD, которые легко интегрируются и демонстрируют высокую точность. Pyannote предлагает комплексный инструмент для анализа аудио, включая детекцию голосовой активности, который также отличается высокой производительностью и простотой интеграции.

2.3 Методы распознавания речи

Методы распознавания речи в последнее время достигают качества, которое сопоставимо и иногда превосходит качество распознавания человеком. Существуют как методы, заточенные под использование в одном домене (например языковом или в условиях определенной зашумленности), так и методы, работающие сразу в нескольких доменах.

Классический метод. Классическая гибридная архитектура [1] для распознавания речи (ASR) включает четыре основные компонента: извлечение акустических признаков из речевых аудиосигналов, акустическое моделирование, языковое моделирование и поиск, основанный на правиле принятия решений Байеса . Акустическое моделирование в этом подходе опирается на скрытые марковские модели, учитывающие вариации скорости речи.

[1] https://github.com/snakers4/silero-vad

[2] https://github.com/pyannote/pyannote-audio

С появлением глубокого обучения в акустическом моделировании глубокие нейросетевые модели заменили HMM и расширили набор акустических признаков, а в языковом моделировании глубокое обучение вытеснило подходы, основанные на подсчете частот. На данный же момент гибридная архитектура распознавания не используется, поскольку появились более эффективные методы, позволяющие строить сквозные системы распознавания речи.

DeepSpeech. Первым подходом к сквозному распознаванию речи является DeepsSpeech [9]. Для данной модели входной аудиосигнал преобразуется в мел-спектрограмму, элементы которой затем подаются в MLP с некоторым шагом. Выходы MLP подаются в двунаправленную рекуррентную нейронную сеть, и для каждого ее скрытого состояния предсказывается распределение вероятностей токенов, из которого потом выбирается наиболее вероятный.

Данная модель показала WER в 16% на датасете Switchboard Hub5'00 [2], при том, что другие модели в то время имели качество не лучше 30%.

DeepSpeech2. DeepSpeech2 [11] является прямым продолжением DeepSpeech1 с усовершенствованной архитектурой. В данной модели вместо MLP используются сверточные нейронные сети, а в качестве рекуррентной нейронной сети используется многослойная сеть GRU. Выходные вектора из рекуррентного блока подаются в однослойный MLP, на выходе которого применяется функция softmax для получения распределения вероятности предсказываемых токенов

Для обучения DeepSpeech2 использовались большие по объему датасеты, чем для DeepSpeech1, кроме того архитектура предполагает лучшую генерализацию, вследствие чего модель достигает WER в 5.33/13.25 на clean/other частях датасета LibriSpeech [12], что равняется качеству распознавания человека.

Wav2Vec. Группа методов wav2vec [25, 29, 21, 28] представляет собой способ предобучения энкодеров для речи в формате самообучения. Обученные энкодеры wav2vec затем используются для построения поверх них акустических моделей, которые выполняют транскрибаацию речи. Качество таких подходов заметно превышает качество акустических моделей, не имеющих предобученных энкодеров.

wav2vec(1.0). Первая версия wav2vec [25] представляет собой полностью сверточную сеть, делящуюся на 2 модуля:

- Encoder – сверточная модель, формирующая на своем выходе низкочастотные представления речи $z_i \in Z$ и кодирующие порядка 30мс сигнала

- Context Network – сверточная модель, принимающая на вход вектора encoder-a $c_i = g(z_i, ..., z_{i-v})$ и кодирующая порядка 210 мс сигнала

В статье используется контрастивный подход обучения, при котором модель учится отличать представление речи z_{i+k}, отстающее на k-шагов от наблюдаемых представлений, от случайных представлений, поставленных на это же место. Функция потерь же имеет вид:

$$L_k = -\sum_{i=1}^{T-k}\Big(log\sigma(z_{i+k}^T h_k(c_i)) + \lambda\mathbb{E}_{\tilde{z}\sim p_n}[log\sigma(-\tilde{z}^T h_k(c_i)]\Big).$$

После обучения модели wav2vec, ее представления c_i используются для обучения акустической модели. В работе показано, что предложенный подход на датасете WSJ [3] превосходит по качеству DeepSpeech2, который до этого был лучшим. Также было показано, что акустические модели, обученные на контекстных представлениях wav2vec, полученных при обучении на небольшом количестве данных (порядка 80 часов), превосходят по качеству модели, обученные на мел-спектрограммах.

wav2vec(2.0). wav2vec2 [29] также состоит из 2 моделей, кроме того, в нем есть модуль, отвечающий за дискретизацию фонем речи:

- Feature Encoder – сверточная модель, формируюшая низкочастотные представления z_i
- Context Network – трансформер, принимающий на вход вектора из encoder-a
- Quantization module – Модуль для отображения представлений z_i в квазифонемы. Латентное представление z_i умножается на матрицу квантизации W_q, в результате получается logit-вектор для данного z_i. Затем logit сэмплируется при помощи Gumbel-Softmax, и получается one-hot вектор (в нем 2 единичных значения – для каждой из codebook групп). Данному вектору сопоставляются вектора из codebooks, и уже вектор с эмбеддингами этих слов умножается на projection matrix W_p для получения итоговых квантизованных векторов.

Функция потерь для обучения складывается из 2 частей:

$$L = L_m + \alpha L_d.$$

L_m – contrastive loss, отвечающий за умение модели отличать истинный квантизованный вектор речи от случайного вектора.

L_d - diversity loss, отвечающий за то, чтобы все псевдофонемы использовались достаточно часто.

$$L_m = -log\frac{exp(sim(c_t q_t)/\kappa)}{\sum_{\tilde{q}\sim Q_t} exp(sim(c_t\tilde{q})/\kappa)}.$$

$$L_d = \frac{1}{GV}\sum_{g=1}^{G} -H(\overline{p}_g) = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V}\overline{p}_{g,v}log(\overline{p}_{g,v}).$$

Wav2vec2 показывает качество, существенно превышающее качество wav2vec1. При этом сами представления аудио получаются настолько качественными, что для достижения метрик WER = 4.8/8.2 на clean/other частях датасета LibriSpeech, достаточно всего 10 минут размеченных аудио.

Whisper. Одной из проблем моделей группы wav2vec является то, что несмотря на предобучение на неразмеченных данных, хорошо описывающих речевой домен, обучение акустических моделей поверх них происходит на довольно специфичных датасетах, вследствие чего качество работы wav2vec на записях, сильно отличающихся от обучающего домена акустических моделей, заметно падает. То есть подход wav2vec предоставляет возможность получать хорошо генерализованные энкодеры речи, однако декодеры оказываются переобучены под конкретный домен.

Авторы модели Whisper [35] предложили создать подход, позволяющий получать как хорошо генерализованный энкодер, так и декодер для речи. Для этого был собран датасет, содержащий 680 тыс. часов размеченных аудио, из которых 117 тыс. на разных языках и 125 тыс. с переводом на английский язык. При этом качество разметки полученного датасета оказалось хуже, чем у общепринятых датасетов для распознавания речи. Для обучения аудио делились на отрезки по 30 с, язык определялся с помощью моделей определения языка, при этом отсекались аудио с низкой степенью уверенности модели в предсказании.

С точки зрения архитектуры Whisper представляет собой классический трансформер. Аудио на входе модели приводится к частоте 16кГц и для него считается мел-спектрограмма. Затем мел-спектрограмма проходит через 2 сверточных слоя, и полученные таким образом векторные представления являются входной последовательностью токенов для трансформера

Хотя основной задачей является распознавания речи, авторами было предложено тренировать модель в многозадачном формате, решая следующие задачи: транскрибация, перевод, детекция интервалов голосовой активности, выравнивание

и идентификация языка. При этом было показано, что модель, обученная в многозадачном формате по метрикам превосходит модели, обученные только распознаванию.

За счет обучения на большом числе различных аудио Whisper хорошо генерализуется на всем домене речи и работает в zero-shot режиме на датасетах, на которых он не обучался. При этом качество Whisper в zero-shot режиме не уступает, а иногда превосходит модели, которые обучались на этих датасетах. По этой причине для задач, в которых домен речи не специализирован, Whisper видится наиболее универсальной моделью среди всех открытых.

RNN-T. Такие модели как wav2vec и Whisper предполагают работу с входными последовательностями речи порядка десятков секунд, то есть их использование в режиме мгновенной транскрибации речи практически невозможна.

RNN-T [23] представляет собой группу методов, которые подходят для работы в реальном времени с минимальной задержкой, когда речь приходит отрезками, длина которых составляет порядка сотен миллисекунд.

Все модели этого типа состоят из 3 блоков:

- Prediction Network – рекуррентная языковая модель, обусловленная на предсказанный в предыдущий момент времени токен y_{u-1}:

$$h_t^{pre} = f^{pre}(y_{u-1}).$$

- Encoder Network – рекуррентная модель, которая кодирует последовательность x_t в представление:

$$h_t^{enc} = f^{enc}(x_t).$$

- Joint Network – полносвязная сеть, комбинирующая h_t^{pre} и h_t^{enc}:

$$z_{t,u} = f^{joint}(h_t^{enc}, h_t^{pre}) = \psi(Uh_t^{enc} + Vh_t^{pre} + b_z).$$

  Распределение вероятности для предсказываемых токенов на выходе Joint Network выглядит следующим образом:

$$P(k|t,u) = softmax(W_y z_{t,u} + b_y).$$

Метрики, получаемые данной группой методов также достигают значений, сравнимых с лучшими из текущих моделей.

Однако использование данных методов видится целесообразным в случае, когда аудиосигнал поступает в систему небольшими интервалами (порядка 100мс), и при этом вычислительных мощностей достаточно для того, чтобы обрабатывать эти интервалы в режиме реального времени. В противном случае, их применение может быть нецелесообразным из-за высоких требований к производительности и сложности настройки.

Онлайн-сервисы. Для распознавания речи практически для каждого языка существуют онлайн-сервисы, предоставляющие API к моделям распознавания речи. Данные сервисы заточены под работу со своим языком и зачастую превосходят по качеству открытые модели.

Для работы с русским языком существуют SaluteSpeech и YandexSpeechKit. Их качество практически сравнимо, однако в условиях сильной защумленности данные сервисы показывают плохое качество, а за счет невозможности предоставления доступа к весам используемых моделей, их применение для таких условий видится затруднительным.

2.4 Методы денойзинга речи

Современные методы для денойзинга речи претерпели значительные изменения с развитием глубокого обучения. Методы, использующие глубокие нейронные сети значительно превзошли традиционные подходы обработки сигналов, за счет использования более сложных архитектур и больших наборов данных для эффективного обучения задаче денойзинга.

Одним из первых подходов является использование глубоких генеративных моделей, таких как вариационные автокодировщики (VAE) [7] и генеративные состязательные сети (GAN) [16]. VAE формируют такие латентные представления сигнала, из которых может быть восстановлена речь без шума. GAN–ы, благодаря своей способности генерировать высококачественные выходные данные, также используются для денойзинга. В GAN-ах одна сеть (генератор) отвечает за создание чистой речи из шумных аудиосигналов, а другая (дискриминатор) отличает чистую речь от очищенной. Эти методы показали хорошие результаты в улучшении качества речи и разборчивости при различных условиях шума.

Сверточные нейронные сети (CNN) [10, 17] и рекуррентные нейронные сети (RNN) также широко используются для денойзинга речи. CNN, благодаря своей спо-

собности захватывать локальные временные и спектральные особенности, эффективно обрабатывают представления кратковременного преобразования Фурье (STFT) речи. RNN, особенно сети с долгой краткосрочной памятью (LSTM) и их варианты, хорошо моделируют временные зависимости в речевых сигналах, что делает их подходящими для последовательных данных, таких как речь. Гибридные модели, сочетающие CNN и RNN, используют сильные стороны обеих архитектур, приводя к улучшению производительности в задаче денойзинга [31].

Значительный прогресс в задача денойзинга показала модель Demucs (Deep Extractor for Music Sources), которая сначала была предложена для задачи разделения музыки [24], а затем адаптирована для денойзинга [26]. Demucs использует автоэнкодерную архитектуру с LSTM-слоями для обработки временного домена сигнала напрямую, а не полагается на представления речи в виде спектрограмм. Этот подход позволяет модели обучаться более хорошим представлениям речи и шума, что приводит к более качественному денойзингу. Архитектура модели предназначена для обработки различных типов шума, что делает ее устойчивой в разных акустических условиях.

Также перспективным методом является использование моделей на основе трансформеров [20], которые показали себя хорошо в задачах обработки естественного языка. Трансформеры, благодаря механизмам внимания, могут захватывать длинные зависимости и контекстную информацию более эффективно, чем RNN. Показано, что трансформерные модели могут обрабатывать всю последовательность речевых сигналов, обучаясь сложным шаблонам разделения шума и речи. Это приводит к значительным улучшениям как в объективных, так и в субъективных оценках качества речи.

Недавние работы также сосредоточены на интеграции нейронных сетей с традиционными методами обработки сигналов [31]. Например, глубокие фильтрационные подходы сочетают нейронные сети с традиционной фильтрацией Винера, где сеть предсказывает параметры фильтра, что улучшает его адаптивность к различным условиям шума. Такие гибридные методы используют сильные стороны как глубокого обучения, так и классической обработки сигналов, достигая передовых результатов в сложных сценариях шумоподавления.

Таким образом, современные нейросетевые методы для денойзинга речи охватывают широкий спектр архитектур и техник, от генеративных моделей, таких как VAE и GAN, до CNN, RNN, трансформеров и гибридных моделей, которые интегри-

руют глубокое обучение с обработкой сигналов. Наиболее сильно выделяется подход Demucs, демонстрируя потенциал глубокого обучения в обработке сигналов во временном домене.

3 Постановка задачи

В работе рассматривается постановка задачи, при которой необходимо в режиме реального времени выделять из сильно зашумленного входящего аудиопотока S интервалы голосовой активности O и производить для них распознавание речи. Данная постановка задачи может быть разбита на 2 отдельные подзадачи:

- Детекция голосовой активности в режиме реального времени
- Распознавание речи в условиях сильной зашумленности

3.1 Детекция голосовой активности в режиме реального времени

В постановке задачи детекции голосовой активности в режиме реального времени предполагается, что имеется последовательность отсчетов аудио $S = (s_1, \ldots, s_T)$, кроме того раз в N отсчетов времени становятся доступны новые отсчеты аудио $I = (s_{T+1}, \ldots, s_{T+N})$. Сама задача состоит в определении того, началась, продолжилась или кончилась голосовая активность в отсчетах I. При этом необходимо выполнение следующих условий:

- Если в интервале I началась или продолжилась голосовая активность, необходимо дождаться следующего интервала
- Если в интервале I кончилась активность, необходимо выделить отсчеты аудио с голосовой активностью O и отправить их на распознавание

Для классификации отсчетов аудио на содержащие и не содержащие голосовую активность используются нейросетевые модели детекции голосовой активности, выполняющие классификацию каждого отсчета аудио на K–классов (в классической постановке K=2):

$$P_S = v(S), P_I = v(I),$$

где $v(\cdot)$ – нейросетевая модель детекции; P_S, P_I – вектора вероятности принадлежности отсчетов к каждому из K–классов размерностей $K \times T$ и $K \times N$.

В качестве класса для каждого отсчета обычно выбирается тот, вероятность которого максимальна:

$$Y_S = \underset{K}{argmax}(P_S), Y_I = \underset{K}{argmax}(P_I).$$

При этом вектора Y_S и Y_I также подвергается процессу постобработки для извлечения из них интервалов, которые задаются отсчетами начала и конца голосовой активности.

Таким образом, задача детекции голосовой активности в реальном времени состоит в определении на основании векторов Y_S и Y_I факта начала, продолжения или окончания голосовой активности в интервале I, и в случае установления факта окончания голосовой активности, выделения для распознавания отсчетов O, содержащих голосовую активность.

3.2 Распознавание речи в условиях сильной зашумленности

В классической постановке задачи распознавания речи предполагается, что имеется последовательность отсчетов аудиосигнала или его мел-спектрограмма $O = (o_1, \ldots, o_T)$, которая в нашем случае получена с помощью детекции голосовой активности в реальном времени.

Задача распознавания заключается в предсказании наиболее вероятной последовательности токенов $\hat{H}$ при условии последовательности O:

$$\hat{H} = \underset{H}{argmax}\, P(H|O),$$

где $P(H|O)$ – условная вероятность последовательности токенов H при условии последовательности O.

Поскольку входная последовательность O является сильно зашумленной, что негативно сказывается на качестве распознавания, в разработанном методе DASR классическая постановка задачи расширяется:

- Последовательность отсчетов O трансформируется в O' при помощи функции денойзинга $f(\cdot) : \mathbb{R}^t \to \mathbb{R}^t$:

$$O' = f(O).$$

- Выходная последовательность $\hat{H}$ трансформируется в последовательность $\hat{H}'$ при помощи функции коррекции транскрибации $s(\cdot) : \mathbb{R}^m \to \mathbb{R}^n$:

$$\hat{H}' = s(\hat{H}).$$

Следует отметить, что функция $f(\cdot)$ предназначена для удаления шума из входной последовательности модели распознавания, а функция $s(\cdot)$ – для коррекции выходной последовательности токенов $\hat{H}$.

Тогда задача распознавания переформулируется в виде:

$$\hat{H}' = s\Big(\underset{H}{argmax}\, P(H|f(O))\Big).$$

Качество распознавания оценивается метриками WER (word error rate) и CER (character error rate), которые сравнивают предсказанную последовательность $\hat{H}$ с истинной:

$$WER = \frac{S + D + I}{S + D + C},$$

где S – число замен, D – число удалений, I – число вставок, C – число корректных слов.

CER считается таким же образом на уровне символов.

4 Метод

4.1 Детекция голосовой активности в режиме реального времени

В условиях реального времени последовательность отсчетов S не является фиксированной, кроме того необходимо для каждого приходящего интервала I устанавливать, начался в нем новый отрезок голосовой активности или продолжается предыдущий. По этой причине невозможно напрямую использовать существующие алгоритмы VAD.

Для адаптации алгоритмов VAD к условиям реального времени предлагается использовать буфер, содержащий в себе предыдущие интервалы, в которых была обнаружена голосовая активность, и на основании совместного анализа буфера и пришедшего интервала устанавливать, какому из следующих случаев соответствует ситуация:

- В интервале началась активность
- В интервале продолжается активность
- В интервале кончилась активность
- В интервале кончилась активность и началась новая

Кроме того, в рассмотрение необходимо принять величину порога, задающего максимальное расстояние между двумя последовательными детектированными интервалами речи. Если расстояние между интервалами больше порога, считается, что это 2 разных интервала речи, иначе считается, что это один интервал.

Тогда для анализа всевозможных наблюдаемых случаев достаточно ввести в рассмотрение 3 условия:

- a – пришедший интервал содержит голосовую активность
- b_1 — предыдущие фрагменты буфера содержат голосовую активность
- b_2 — дистанция между началом голосовой активности пришедшего интервала и концом голосовой активности буфера меньше заданного порога

Таблица 1: Действия для выражений

Условие	Случай	Действие
$a \wedge \neg b_1$	I	Буфер дополняется пришедшим интервалом
$a \wedge b_1$	II	Буфер дополняется пришедшим интервалом
$\neg a \wedge b_1 \wedge b_2$	III	Элементы буфера с голосовой активностью объединяются в аудио, затем буфер обновляется
$a \wedge b_1 \wedge \neg b_2$	IV	Элементы буфера с голосовой активностью объединяются в аудио, затем буфер обновляется, и в него добавляется пришедший интервал

В таблице 1 представлены соответствующие наборы условий для каждого случая, а также указаны действия, которые необходимо выполнить с буфером и входящим интервалом. На основании данных условий представлен псевдокод разработанного алгоритма 4.1 детекции голосовой активности в реальном времени – RealVADR.

Алгоритм 4.1 Алгоритм обработки интервалов звука RealVADR

Вход: buffer, new_interval, threshold

Выход: updated_buffer

a := v(buffer, new_interval) ≠ None
b_1 := buffer ≠ None
b_2 := (max_right_bound(buffer) - max_left_bound(new_interval)) < threshold
если a **то**
    Добавить интервал в буфер
    если b_1 & $\neg b_2$ **то**
        Объединить элементы буфера с голосовой активностью в аудио
        Отправить аудио на распознавание
        Обновить буфер
        Добавить интервал в буфер
    иначе
        Продолжить обработку следующих интервалов
иначе
    если b_1 & b_2 **то**
        Объединить элементы буфера с голосовой активностью в аудио
        Отправить аудио на распознавание
        Обновить буфер
    иначе
        Продолжить обработку следующих интервалов

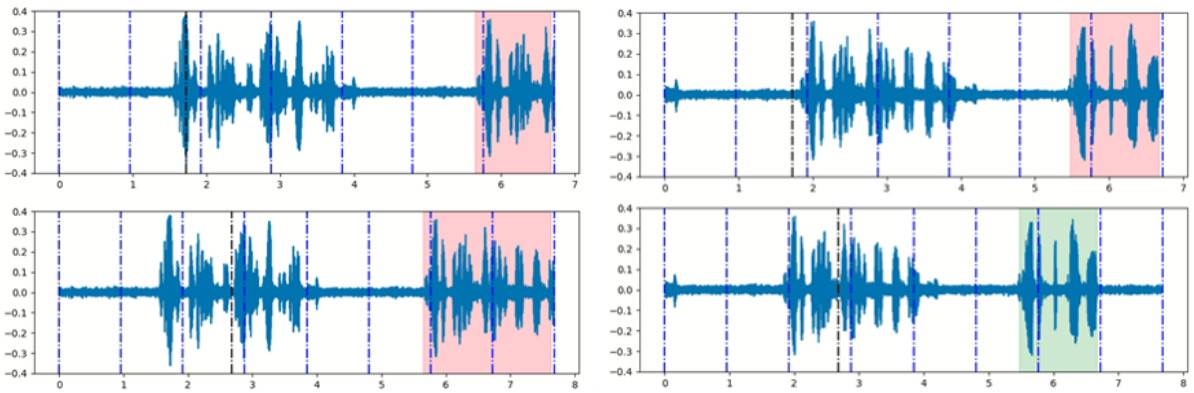

Рис. 1: Иллюстрация работы алгоритма RealVADR в реальном времени

Предложенный алгоритм RealVADR оперирует всего двумя параметрами: размером входящего звукового интервала и пороговым значением. Для правильной на-

стройки порога важно учитывать естественные характеристики конкретного языка, на котором будет применяться данный алгоритм.

При определении размера интервала следует учесть, что алгоритм RealVADR выполняет прогнозирование после завершения каждого интервала, поэтому его продолжительность должна быть не меньше времени, необходимого для работы базовой модели VAD. С другой стороны, размер интервала также влияет на общую задержку алгоритма, поэтому важно найти баланс между размером интервала и количеством предсказаний, выполняемых моделью.

4.2 Распознавание речи в условиях сильной зашумленности

4.2.1 Денойзинг входной последовательности

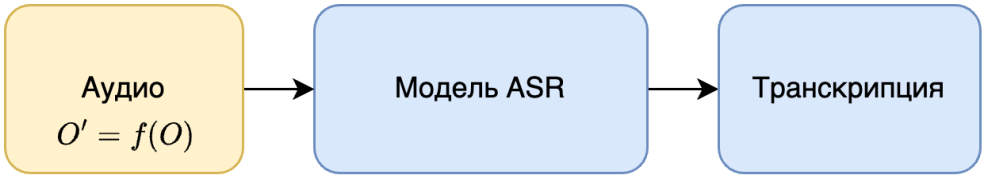

Рис. 2: Денойзинг входной последовательности в распознавании речи

В данной части разработанного метода DASR предполагается улучшение качества транскрибации путем очистки входного сигнала от присутсвующего в нем шума. Для этого предлагается использовать нейросетевые модели денойзинга [26], удаляющие из входного сигнала шум (как показано на рисунке 3).

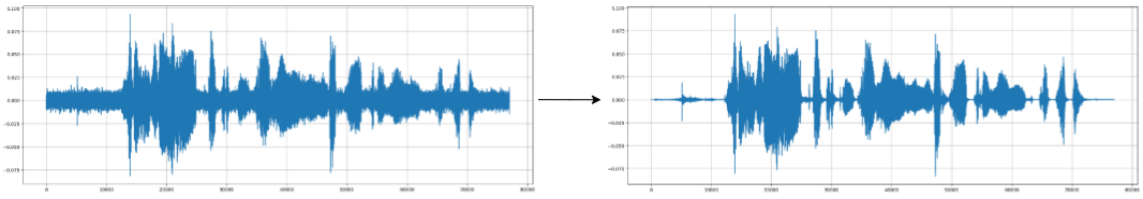

Рис. 3: Пример удаление шума из входной последовательности

В классической постановке задачи денойзинга решается задача построения функции расшумления $f(\cdot)$ для аудиосигнала O, содержащего чистый сигнал Y и аддитивный шум $\mathcal{N}$:

$$f(O) = f(Y + \mathcal{N}) \approx Y.$$

В предложенной в данной работе постановке функция $f(\cdot)$ меняет свое назначение, и служит для изменения входной последовательности O таким образом, чтобы

максимизировать вероятность предсказываемой моделью распознавания последовательности токенов $\hat{H}$:

$$\hat{H} = \underset{H}{argmax}\, P(H|f(O)).$$

Такая постановка задачи позволяет использовать для обучения денойзеров CTC-функцию потерь, которая используется для обучения моделей распознавания и позволяет оптимизировать метрики WER и CER. При этом в данной постановке не требуется, чтобы $f(O) \approx Y$.

Следует отметить, что для обучения денойзеров предполагается использование моделей распознавания, относительно выходов которых будет рассчитываться CTC-функция потерь, и через которую будут проходить градиенты для обучения денойзеров. При этом предложенный метод DASR предполагает, что веса модели распознавания заморожены, а улучшение качества распознавания достигается только за счет обучения денойзера.

В качестве модели–денойзера в работе используется автоэнкодер архитектуры Demucs [24]. Данная модель состоит из сверточных энкодера и декодера, между слоями которых присутствуют skip connection. Кроме того, между энкодером и декодером находится рекуррентная нейронная сеть LSTM, необходимая для сохранения информации при обработке последовательных интервалов речи. Архитектура Demucs представлена на рисунке 4.

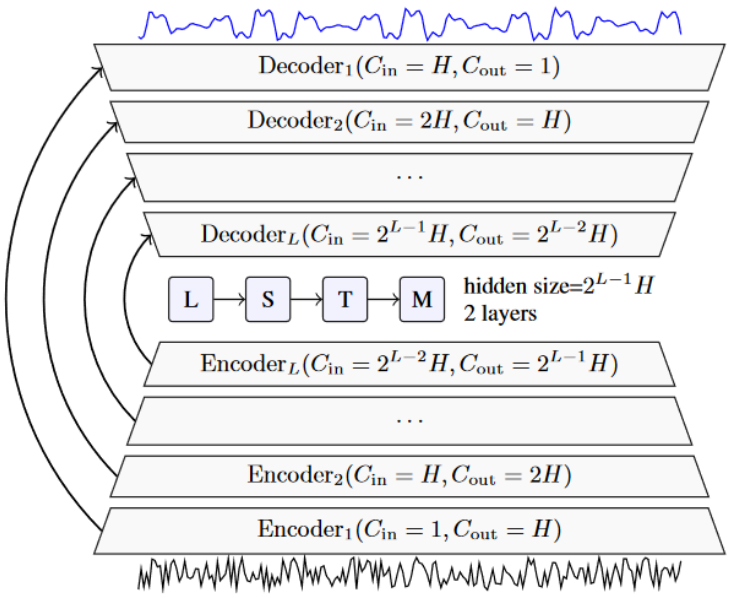

Рис. 4: Архитектура Demucs

Перед обучением Demucs инициализируется весами модели facebook denoiser [26],

которая обучалась задаче удаления шума из речи. При этом несмотря на другое назначение модели facebook denoiser, ее веса являются хорошей стартовой точкой для обучения, поскольку постановка задачи удаления шума хорошо коррелирует с нашей постановкой задачи.

4.2.2 Коррекция выходной транскрибации

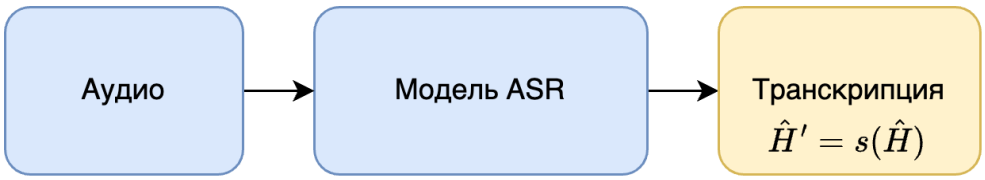

Рис. 5: Коррекция выходной транскрибации для распознавании речи

В данной части разработанного метода DASR предполагается улучшение качества распознавания путем коррекции выходной транскрибации моделей распознавания речи. Данная часть метода оправдана, поскольку модели распознавания не являются полноценными языковыми моделями, и их транскрибации могут содержать неточности и опечатки, которые возможно исправить (как показано на рисунке 6).

Повышение качества распознавания достигается за счет использования моделей коррекции выходной последовательности $s(\cdot)$, преобразующих транскрибации моделей распознавания:

$$\hat{H}' = s(\hat{H}).$$

В качестве таких моделей могут использоваться такие seq2seq языковые модели, как например T5 [27] или FRED-T5 [36], а также инструктивные большие языковые модели [37, 38], для которых подбирается соответствующий задаче коррекции транскрибации промпт.

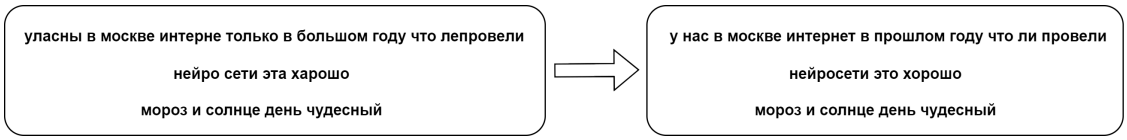

Рис. 6: Пример коррекции выходной транскрибации

5 Вычислительные эксперименты

Поскольку предложенный метод состоит из 2 частей – алгоритма детекции голосовой активности в реальном времени RealVADR и метода распознавания речи в условиях сильной зашумленности DASR, вычислительные эксперименты для удобства приведены для каждой из частей отдельно.

5.1 Детекция голосовой активности в режиме реального времени

5.1.1 Входные данные

Так как предложенный метод RealVADR используется для детекции голосовой активности в реальном времени с последующим распознаванием речи, его качество следует оценивать исходя из качества на датасетах для распознавания речи.

Для тестирования алгоритма RealVADR был создан датасет, записанный на роботе Husky в условиях шума, создаваемого роботом. Датасет состоит из коротких команд для робота и имеет 4 части:

- clear_voice – команды, записанные в условиях без шума
- noise_voice_1m – команды, записанные в условиях шума, на расстоянии 1м от робота
- noise_voice_2m – команды записанные в условиях шума на расстоянии 2м от робота
- road_voice – команды, записанные на фоне едущего робота

Кроме того, для тестирования предложенного алгоритма используется тестовая выборка русскоязычной части датасета CommonVoice [22].

5.1.2 Анализ результатов

Влияние VAD на качество распознавания. Было замечено, что при использовании результатов VAD в качестве входных данных для моделей распознавания напрямую, качество оказывается хуже, чем при использовании целых аудио из датасета.

Следует отметить, что результаты VAD обрадают следующей особенностью: детектированные интервалы начинаются/заканчиваются ровно в моменты начала/окончания речи и не имеют отрезков без речи. С точки зрения распознавания это означает, что для модели недоступны отрезки, показывающие уровень шума в среде без голосовой активности, что скорее всего и приводит к снижению качества, то есть между выходным распределением данных VAD и входным распределением данных моделей ASR имеется сдвиг.

В связи с этим логично добавить к результатам VAD зазор, не содержащий голосовой активности, как это показано на рисунке 7. Такое преобразование может позволить устранить сдвиг в данных и улучшить качество распознавания.

Для экспериментов по оценке влияния величины зазора на качество распознавания было рассмотрено 4 модели: Whisper-large-v2, wav2vec2-xls-r-1b-russian, а также сервисы распознавания речи YandexSpeechKit и SaluteSpeech.

На рисунке 8 показано, как зависят метрики CER и WER рассмотренных моделей для выбранных датасетов от величины зазора. Разными цветами показаны графики для рассмотренных моделей, а пунктирными линиями — офлайн-метрики моделей.

Для всех случаев характерно значительное улучшение метрик по сравнению с нулевым значением зазора (улучшение достигает порядка 10-15 единиц WER и CER), что говорит о возможности улучшения метрик путем введения зазора и устранении сдвига распределения между выходными данными модели VAD и входными данными моделей ASR.

Для наименее зашумленных датасетов clear_voice и CommonVoice метрики CER и WER улучшаются при увеличении зазора. При достижении зазором значения 0.5, метрики большинства моделей распознавания сравниваются с офлайн-метриками, а в некоторых случаях и незначительно превосходят их.

Для более зашумленных частей датасета такая зависимость также прослежива-

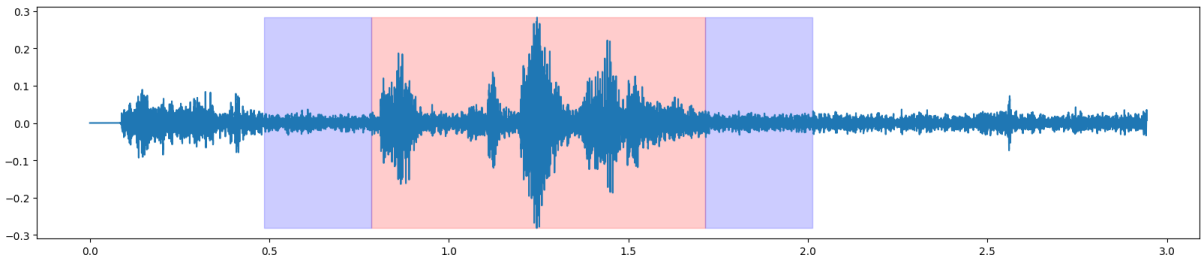

Рис. 7: Детектированнный звук с добавленным зазором

Таблица 2: Сравнение метрик качества WER/CER методов на различных наборах данных

Метод	noise_voice_1m	noise_voice_2m	road_voice	clear_voice	commonvoice
wav2vec2 (offline)	32.60 / **11.07**	55.84 / **24.36**	80.48 / **38.05**	**8.49** / **1.12**	**15.01** / **4.74**
RealVADR(wav2vec2)	32.60 / 12.82	**51.94** / 25.75	**70.73** / 40.21	11.32 /2.10	16.40 / 6.19
Whisper (offline)	39.30 / 23.55	64.93/ 39.91	74.39 / 52.55	23.58 / 17.39	**12.41** / **5.48**
RealVADR(Whisper)	**38.04** / **22.67**	**61.03** / **37.12**	**67.07** / **45.54**	**20.75** / **15.84**	13.75 / 6.69
YandexSpeechKit (offline)	**32.22**/ **15.76**	**32. 43**/ **20.55**	**55.55** / 47.47	4.12 / 1.34	/
RealVADR(YandexSpeech)	34.44 / 20.12	47.29 / 29.72	58.02 / **46.71**	4.12 / 1.34	/
SaluteSpeech (offline)	**28.26** / 16.34	**41.55** / **24.12**	**46.34** / **28.29**	**3.77** / **1.12**	**10.15** / **4.89**
RealVADR(SaluteSpeech)	29.34 / **15.64**	44.15 / 26.91	50.00 / 36.81	5.66 / 2.66	11.59 / 6.37

ется, однако при увеличении зазора выше 0.2с наблюдается флуктуация метрик и в некоторых случаях, например в случае с Whisper и YandexSpeechKit, они перестают улучшаться.

Опираясь на построенные графики, можно взять значение зазора 0.2с за оптимальное и зафиксировать. В таком случае можно сравнить офлайн-метрики и значения метрик для всех алгоритмов при выбранном значении зазора. Данное сравнение приведено в таблице 2.

Можно отметить, что значения метрик для данной величины зазора незначительно хуже офлайн метрик (ухудшение составляет в среднем 1-5 единиц WER или CER), то есть итоговая система распознавания речи в реальном времени, построенная на основе предложенного метода RealVADR, будет близка по качеству к офлайн-распознаванию.

5.1.3 Выводы

Анализ результатов проведенных экспериментальных исследований позволяет сделать следующий выводы:

- Возможно адаптирование существующих моделей VAD для задачи детекции голосовой активности в реальном времени
- Для адаптации моделей VAD к детекции в реальном времени разработан метод RealVADR, имеющий всего 2 параметра – размер входящего интервала речи и порог, задающий максимальное время без речи
- Регулирование зазора при обработке результатов детекции голосовой активности позволяет значительно улучшить качество распознавания и приблизить его к качеству офлайн-распознавания с разницей в 1-5 единиц WER

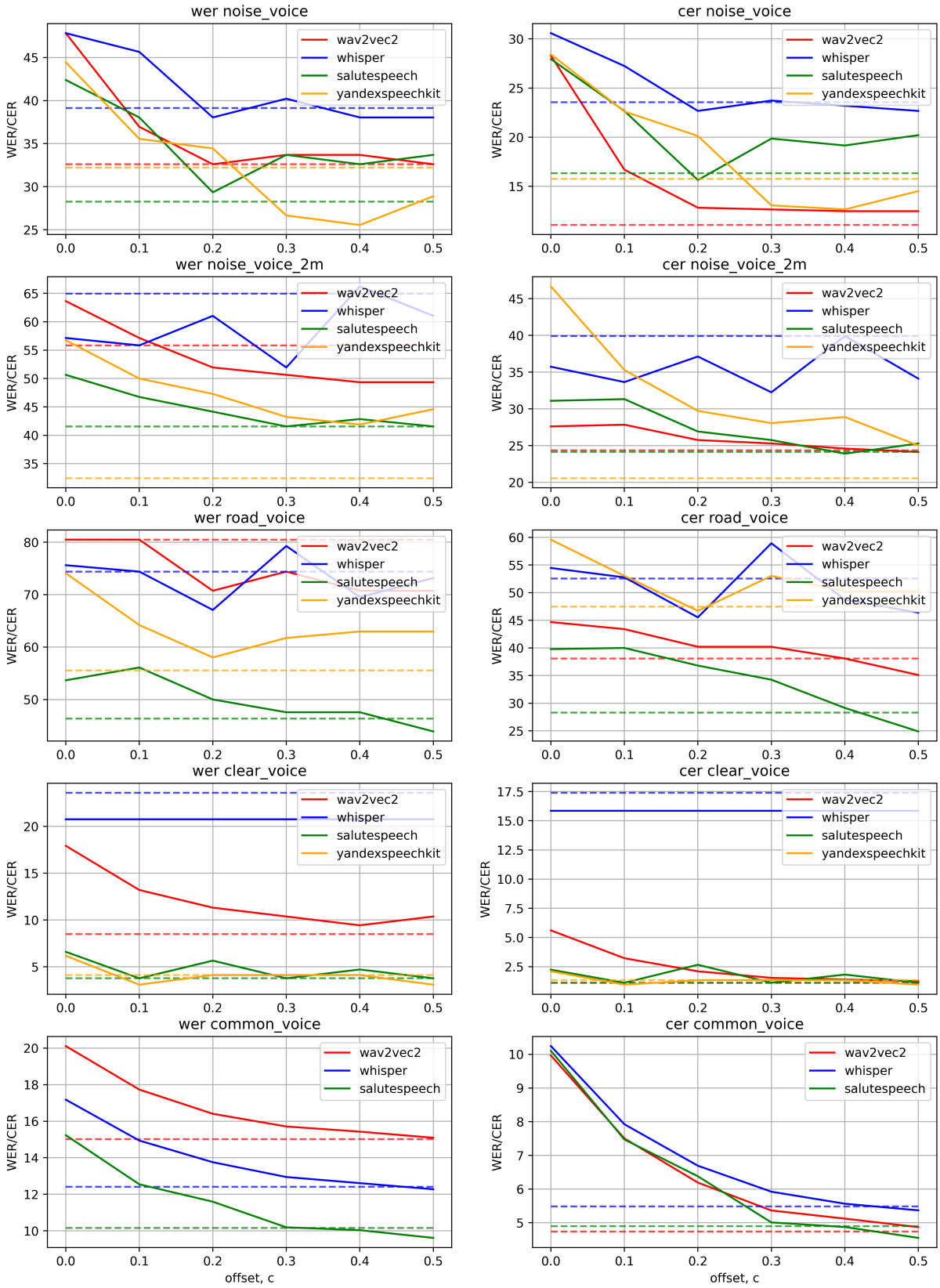

Рис. 8: Зависимость метрик распознавания речи WER и CER от величины зазора, используемой в подходе RealVADR

Алгоритм 5.1 Алгоритм добавления шума с заданным SNR

Вход: signal, noise, SNR

Выход: noisy_signal

1: $P_{signal} := \sum_{i=1}^{N} signal[i]^2$ — Определение мощности исходного сигнала

2: $P_{noise} := \sum_{i=1}^{M} noise[i]^2$ — Определение мощности шума

3: $noise := \sqrt{\frac{P_{signal}}{P_{noise}}} \cdot noise$ — Выравнивание сигналов по мощности

4: $noisy_signal := \sqrt{SNR} \cdot signal + noise$ — Создание зашумленного сигнала с желаемым SNR

5.2 Распознавание речи в условиях сильной зашумленности

5.2.1 Входные данные

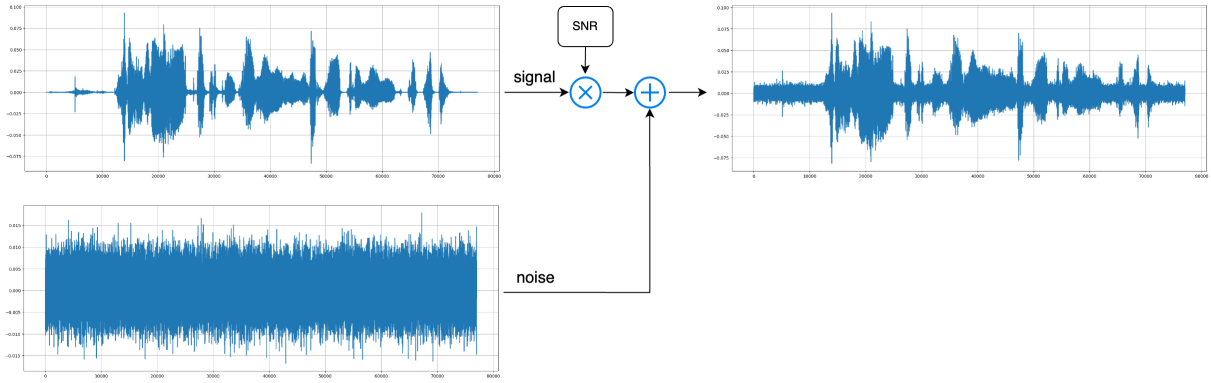

Рис. 9: Формирование зашумленных аудио с желаемым уровнем SNR

Для решения задачи распознавания речи в условиях сильной зашумленности предложена следующая методика формирования обучающих и тестовых датасетов:

- Берутся аудио, не содержащие шума или содержащие пренебрежимо малый шум
- Формируется случайный белый шум
- Фиксируется желаемый уровень SNR (signal to noise ratio)
- Из оригинального аудио и сгенерированного белого шума формируется новое зашумленное аудио с желаемым уровнем SNR согласно алгоритму 5.1. Также формирование зашумленного аудио показано на рисунке 9

В качестве датасетов, из которых берутся аудио, не содержащие шума рассматриваются русскоязычная и англоязычная часть CommonVoice [22], русскоязычные

датасеты SOVA RuDevices [33] и SberGolos [30] и англоязычный датасет LibriSpeech [12].

5.2.2 Дообучение моделей распознавания

Базовым подходом, с которым сравнивается предложенный метод является классическое дообучение моделей распознавания. Данный подход гарантированно приводит к улучшению качества моделей распознавания, кроме того, качество дообученных моделей распознавания в большинстве случаев можно рассматривать как максимально достижимое в задаче качество. Однако следует отметить, что дообученные модели теряют в генерализации на всем множестве аудио.

Для экспериментов по дообучению используется следующая экспериментальная постановка:

- В качестве базового датасета используются русскоязычный и англоязычный CommonVoice. Согласно 5.2.1 формируются тренировочный и тестовый датасет с уровнем SNR=10.
- В качестве моделей распознавания используются модели wav2vec для русского[3] и английского языка[4]
- Производится дообучение моделей распознавания с использованием CTC-функции потерь

Результаты экспериментов по дообучения моделей распознавания представлены в таблице 3.

Таблица 3: Эксперименты по дообучению моделей распознавания речи на CommonVoice

Метод	WER	CER
wav2vec(def)-russian	43.61	15.92
wav2vec(tune)-russian	**30.53**	**9.32**
wav2vec(def)-english	40.11	20.38
wav2vec(tune)-english	**34.86**	**16.81**

[3] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-russian

[4] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english

Как видно из экспериментов, для обоих рассмотренных языков дообучение приводит к улучшению метрик распознавания.

5.2.3 Использование недообученных денойзеров

Поскольку качество распознавания в рассматриваемых экспериментах ухудшается только за счет добавления шума, естественным базовым решением видится использование денойзеров, которые обучены задаче удаления шума из сигнала. Для таких экспериментов используются следующая экспериментальная постановка:

- Используется тестовая часть русскоязычного датасета CommonVoice и зашумляется белым шумом с SNR=10
- Используется модель распознавания wav2vec для русского языка
- Используется оригинальная модель facebook denoiser [26]
- Производится денойзинг зашумленного датасета и для него выполняется транскрибация
- Также производится транскрибация зашумленного датасета, но без применения денойзинга

Результаты данных экспериментов приведены в таблице 4.

Таблица 4: Результаты экспериментов с недообученным денойзером

Метод	WER	CER
wav2vec(def)	43.61	15.92
wav2vec(def)+denoiser(def)	47.92	19.33

Из экспериментов видно, что использование недообученного денойзера ухудшает качество по метрикам WER и CER примерно на 4 единицы по сравнению с результатами модели распознавания без использования денойзинга. Данное явление объясняется тем, что денойзер сильно изменяет аудиосигнал, в результате чего он становится сильно отличным от входного распределения модели распознавания, поэтому, несмотря на удаление шума из сигнала, итоговое качество распознавания оказывается хуже. Различия между оригинальным сигналом и сигналом, подверженным денойзингу, хорошо видны на мел-спектрограммах и показаны на рисунке 10, из которого видно, что денойзером удаляются целые частотные полосы.

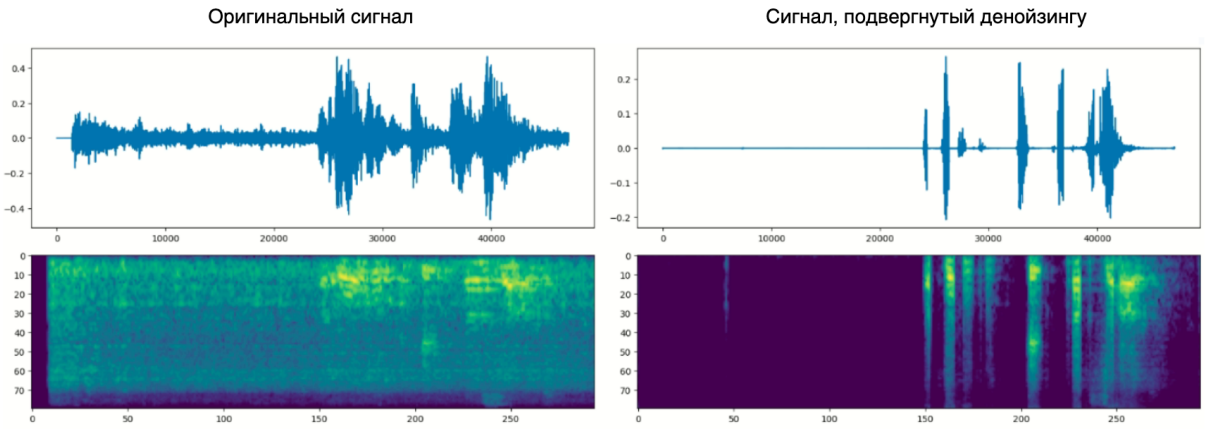

Рис. 10: Мел-спектрограммы оригинального сигнала и сигнала, подвергнотому денойзингу

5.2.4 Дообучение денойзеров

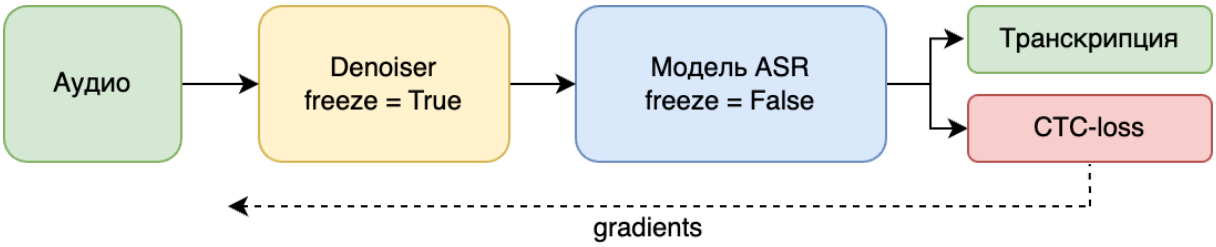

Рис. 11: Схема обучения денойзера на выходах модели распознавания в методе DASR

Для дообучения денойзеров методом DASR используется следующая постановка экспериментов:

- В качестве датасетов для обучения и валидации используются русскоязычный и англоязычный CommonVoice, зашумленные белым шумом с SNR=10
- В качестве денойзера используется facebook denoiser с предобученными весами. При этом рассматривается 2 случая – когда энкодер модели заморожен (то есть его веса при обучении не изменяются), и когда энкодер модели разморожен
- В качестве модели распознавания используется wav2vec для русского или английского языка
- Веса модели wav2vec замораживаются, и в процессе обучения не изменяются
- В процессе обучения аудио сначала подвергается денойзингу, после чего транскрибации, и на выходе модели распознавания рассчитывается CTC-функия потерь. Путем обратного распространения ошибки через модель распознавания

считаются градиенты для денойзера, и происходит обновление его весов, как показано на рисунке 11. При этом веса модели распознавания не изменяются.

Согласно данной постановке было проведено 2 группы экспериментов по дообучению денойзеров для русского и английского языков, их результаты представлены в таблицах 5 и 6 соответственно.

Таблица 5: Эксперименты по дообучению денойзеров на русском языке

Метод	WER	CER
wav2vec(def)	43.61	15.92
wav2vec(def)+denoiser(fine)	**30.94**	**9.87**
wav2vec(def)+denoiser(fine, freeze encoder)	31.65	10.22
wav2vec(fine)	**30.53**	**9.32**

Таблица 6: Эксперименты по дообучению денойзеров на английском языке

Метод	WER	CER
wav2vec(def)	40.11	20.38
wav2vec(def)+denoiser(fine)	**34.76**	16.90
wav2vec(def)+denoiser(fine, freeze encoder)	34.88	17.15
wav2vec(fine)	34.86	**16.81**

Из экспериментов видно, что для русского языка качество предложенного метода DASR практически сравнимо с результатами дообучения, а для английского языка даже превосходит его. Исходя из этого можно сделать вывод, что предложенный метод DASR позволяет достигать качества распознавания, сравнимого с максимально достижимым, и при этом нет необходимости изменения весов модели распознавания.

5.2.5 Использование непредобученных денойзеров

Поскольку задача денойзинга несколько отличается от предложенной постановки, и использование оригинальных денойзеров без дообучения не приводит к улучшению качества распознавания, как показано в 5.2.3, рассматриваются эксперименты по обучению денойзеров с нуля.

Для таких экспериментов используется следующая экспериментальная постановка:

- В качестве датасета для обучения и валидации используется русскоязычный CommonVoice, зашумленный белым шумом с SNR=10
- Используется модель распознавания wav2vec для русского языка
- В качестве денойзера используется модель архитектуры Demucs [24], инициализированная случайно
- Производится денойзинг зашумленного датасета и для него выполняется транскрибация
- Производится обучение денойзера аналогично 5.2.4

На рисунке 12 представлены графики метрик WER, CER и CTC-функции потерь на валидационном датасете для случайно инициализированного и предобученного денойзеров. Как видно из графиков, метрики и функция потерь для случайно инициализированного денойзера оказываются сильно хуже, чем для предобученного.

Из этого можно сделать вывод, что несмотря на отличающуюся от предложенной постановку задачи, предобученные денойзеры являются хорошей стартовой точкой для обучения и содержат в себе релевантную информацию, вследствие чего их использование позволяет получать при обучении существенно лучшее качество распознавания.

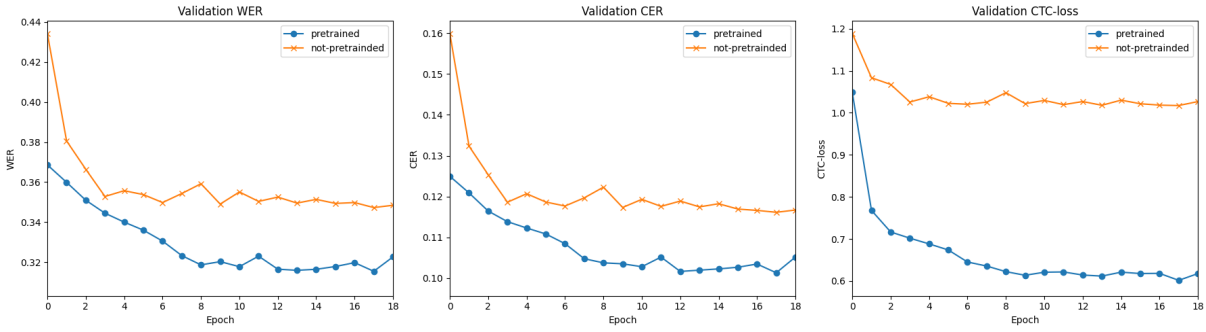

Рис. 12: Метрики и функция потерь на валидационном датасете для предобученного и случайно инициализированного денойзеров

5.2.6 Дообучение денойзеров с использованием дополнительной функции потерь

Поскольку веса денойзеров хорошо подходят для начальной инициализации при обучении, и задача денойзинга близка к рассматриваемой постановке задачи, есте-

ственно выглядит интегрировать в функцию потерь для обучения дополнительную функцию потерь из оригинальной постановки задачи денойзинга.

Данная функция потерь содержит 2 составляющие: l_1–норму между оригинальным сигналом y и восстановленным сигналом $\hat{y}$, а также разницу между спектральными составляющими исходного и восстановленного сигнала L_{sfft}:

$$L_{denoiser} = \frac{1}{T}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_1 + \sum_{i=1}^{M} L_{\text{stft}}^{(i)}(\boldsymbol{y}, \hat{\boldsymbol{y}}),$$

$$L_{\text{stft}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = L_{sc}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + L_{\text{mag}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right),$$

$$L_{sc}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\| |\operatorname{STFT}(\boldsymbol{y})| - |\operatorname{STFT}(\hat{\boldsymbol{y}})| \|_F}{\| |\operatorname{STFT}(\boldsymbol{y})| \|_F},$$

$$L_{\text{mag}}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = \frac{1}{T}\| \log |\operatorname{STFT}(\boldsymbol{y})| - \log |\operatorname{STFT}(\hat{\boldsymbol{y}})| \|_1,$$

где STFT (Short-Time Fourier Transform) – кратковременное преобразование Фурье.

С учетом функции потерь денойзинга, итоговая функция потерь для обучения будет складываться из 2 частей:

$$L = L_{CTC} + \alpha L_{denoiser},$$

где α – коэффициент, задающий степень значимости функции потерь денойзинга. В экспериментах данный коэффициент принимает значение 0.01.

Для экспериментов по добавлению дополнительной функции потерь использовалась та же постановка, что и в 5.2.4. Результаты данных экспериментов представлены в таблице 7 (RU означает русскоязычный wav2vec и русскоязычную часть датасета CommonVoice, а EN – англоязычные).

Таблица 7: Эксперименты по добавлению дополнительной функции потерь денойзинга

Метод	Язык	Функция потерь	WER	CER
wav2vec+denoiser	RU	L_{CTC}	**30.94**	**9.32**
wav2vec+denoiser	RU	$L_{CTC} + 0.01L_{denoiser}$	33.45	10.96
wav2vec+denoiser	EN	L_{CTC}	**34.76**	**16.90**
wav2vec+denoiser	EN	$L_{CTC} + 0.01L_{denoiser}$	36.36	18.02

Как видно из таблицы, добавление функции потерь денойзинга ухудшает качество распознавания. Это подтверждает, что не смотря на схожую постановку задачи, для задачи распознавания нет необходимости строго соответствовать исходной постановке задачи денойзинга.

5.2.7 Коррекция транскрибации

Для экспериментов по коррекции транскрибации предложенным методом DASR используется следующая экспериментальная постановка:

- Используется тестовая часть русскоязычного датасета CommonVoice, которая зашумлена белым шумом с SNR=10
- Тестовый датасет транскрибируется с помощью моделей распознавания (в том числе и моделями с денойзингом)
- Полученные транскрибации корректируются с помощью модели T5[5], а также с помощью YandexSpeller

Результаты данных экспериментов представлены в таблице 8.

Таблица 8: Эксперименты по коррекции транскрибации

Метод	WER	CER
wav2vec(def)	43.61	15.92
wav2vec(def)+ruT5	36.57	19.43
wav2vec(def)+YandexSpeller	**36.20**	**14.90**
wav2vec(def) + denoiser(fine)	30.94	9.87
wav2vec(def) + denoiser(fine) + ruT5	**24.32**	11.69
wav2vec(def) + denoiser(fine) + YandexSpeller	24.39	**8.92**
wav2vec(fine)	30.53	9.32
wav2vec(fine) + ruT5	**23.61**	11.32
wav2vec(fine) + YandexSpeller	24.35	**8.54**

Как видно из результатов экспериментов, коррекция транскрибации во всех случаях положительно влияет на метрику WER и снижает ее на 6-8 единиц, однако при

[5] https://huggingface.co/bond005/ruT5-ASR

этом в случае с ruT5 приводит к ухудшению метрики CER на 2-4 единицы. Данное явление объясняется «галлюцинированием» языковых моделей и заменой некоторых слов на синонимы, а также добавлением несуществующих, но подходящих по контексту слов, что иллюстрируется рисунком 13. Для YandexSpeller такого явления не наблюдается, поскольку он не является полноценной языковой моделью, и помимо снижения WER он позволяет снизить CER примерно на одну единицу во всех случаях.

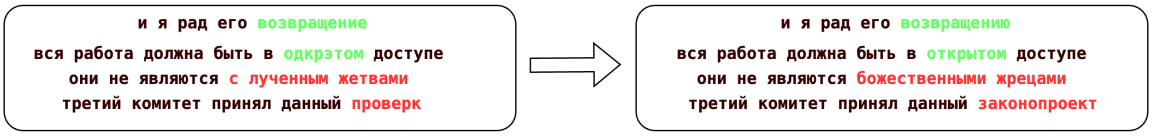

Рис. 13: Примеры коррекции транскрибации с помощью ruT5

Поскольку наиболее важной метрикой для распознавания является WER, то в случае с ruT5 на фоне ухудшения CER на 2-4 единицы, улучшение WER на 6-8 единиц является улучшением.

Следует также отметить, что коррекция транскрибации в сочетании с деноизингом входного сигнала практически вдвое снижает WER по сравнению с исходной моделью, что демонстрирует существенное улучшение. Кроме того, примечательно, что коррекция транскрибации показывает сравнимые метрики как при применении к дообученной модели распознавания, так и к модели с примененным денойзингом.

5.2.8 Генерализация денойзинга на других датасетах

Поскольку важным критерием для моделей распознавания является возможность их генерализации на другие речевые домены, обученные на CommonVoice денойзеры тестируются на других датасетах, на которых они не обучались. В качестве датасетов для тестирования были выбраны: мультиязычный Google Fleurs [34], русскоязычные SOVA [33] и SberDevices Golos [30], а также англоязычный датасет LibriSpeech [12]. Для тестирования берутся только тестовые части выбранных датасетов.

Следует отметить, для англоязычных датасетов используются обученные на англоязычном CommonVoice денойзеры и базовые модели wav2vec, а для русскоязычных используются русскоязычные wav2vec и денойзер. Результаты тестирования приведены в таблице 9.

Таблица 9: Генерализация денойзеров на других датасетах

Тестовый датасет	Модель распознавания	Денойзер	WER	CER
Google Fleurs (RU)	wav2vec(RU)	–	61.59	24.90
	wav2vec(RU-fine)	–	44.90	14.83
	wav2vec(RU)	+	**43.66**	**14.67**
SOVA RuDevices	wav2vec(RU)	–	73.61	35.83
	wav2vec(RU-fine)	–	**63.48**	**27.61**
	wav2vec(RU)	+	64.23	28.53
SberDevices Golos (farfield)	wav2vec(RU)	–	96.18	42.69
	wav2vec(RU-fine)	–	**82.82**	**33.80**
	wav2vec(RU)	+	90.21	38.25
SberDevices Golos (crowd)	wav2vec(RU)	–	76.84	30.05
	wav2vec(RU-fine)	–	**65.21**	**22.35**
	wav2vec(RU)	+	66.93	23.53
Google Fleurs (EN)	wav2vec(EN)	–	32.56	15.53
	wav2vec(EN-fine)	–	27.81	12.63
	wav2vec(EN)	+	**25.67**	**11.91**
LibriSpeech Clean (EN)	wav2vec(EN)	–	14.21	5.00
	wav2vec(EN-fine)	–	12.19	4.03
	wav2vec(EN)	+	**11.05**	**3.74**
LibriSpeech Other (EN)	wav2vec(EN)	–	29.42	13.21
	wav2vec(EN-fine)	–	25.13	10.56
	wav2vec(EN)	+	**23.50**	**9.98**

Как видно из таблицы, на всех датасетах использование денойзеров улучшает качество распознавания, причем в случае с русскоязычным Google Fleurs, улучшение WER составляет 18 единиц по сравнению с моделью без денойзера. В сравнении с дообученными на CommonVoice wav2vec, денойзеры в случае SOVA и SberDevices Golos показывают сопоставимое качество, а в случае Fleurs и LibriSpeech превосходят дообученные wav2vec.

Таким образом, денойзеры, обученные с помощью разработанного метода DASR, показывают хороший уровень генерализации, и в большинстве случаев их генерализация оказывается лучше, чем у способа дообучения. Данный факт связан с тем, что

для денойзинга меняется только сам аудиосигнал, то есть низкоуровневое представление речи, тогда как в случае с дообучением меняются веса модели распознавания, которые оперируют уже высокоуровневыми представлениями и более подвержены переобучению.

5.2.9 Генерализация денойзинга на других моделях

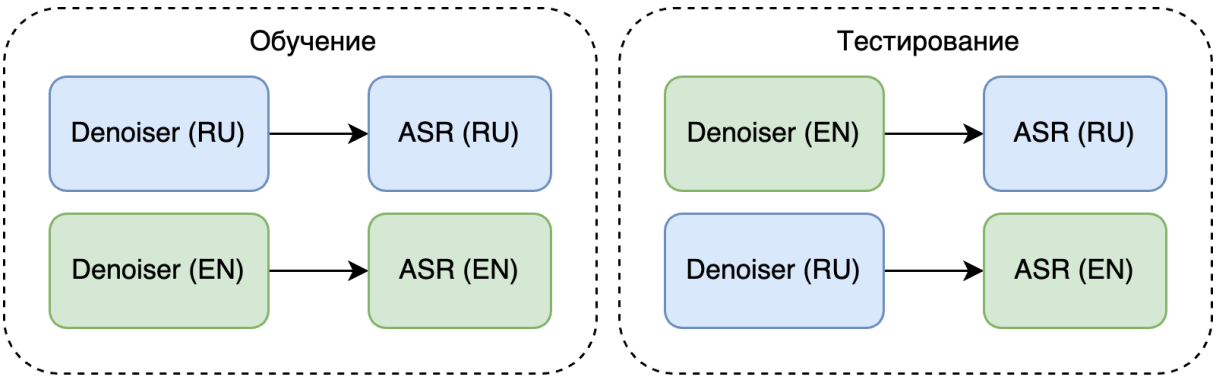

Рис. 14: Генерализация денойзеров, обученных методом DASR, между разными моделями

Так как денойзинг является сильно регуляризованным методом и изменяет только входной аудиосигнал модели распознавания, следует рассмотреть возможность генерализации денойзинга не только на другие датасеты, но и на другие модели.

Для таких экспериментов используется следующая постановка (также она иллюстрируется рисунком 14):

- Для русскоязычного wav2vec используется денойзер, который получен при обучении на англоязычном CommonVoice с англоязычным wav2vec.
- Для англоязычного wav2vec используется денойзер, который получен при обучении на русскоязычном CommonVoice с русскоязычным wav2vec
- Полученные модели тестируются на тестовых частях русскоязычного и англоязычного CommonVoice

Как видно из экспериментов, использование денойзеров с теми моделями распознавания, на выходах которых они не обучались, приводит к увеличению качества. При этом отличие между использованием оригинального и неоригинального денойзера составляет порядка 2 единиц WER.

Таблица 10: Генерализация денойзеров между разными моделями

Модель распознавания	Тестовый датасет	Денойзер	WER	CER
wav2vec(EN)	CommonVoice(EN)	–	40.11	20.38
		denoiser (EN)	34.76	16.90
		denoiser (RU) – zero-shot	36.25	18.20
wav2vec(RU)	CommonVoice(RU)	–	43.61	15.92
		denoiser (RU)	30.94	10.22
		denoiser (EN) – zero-shot	32.41	10.51

Таким образом, денойзеры, получаемые в результате обучения предложенным методом DASR, позволяют достичь такого уровня генерализации, при котором возможно их использование не только с теми моделями распознавания, на выходах которых они обучались, но и для других моделей распознавания. Данный факт свидетельствует о том, что денойзеры не переобучаются под конкретную модель распознавания, а изменяют сигнал таким образом, чтобы его можно было распознать более качественно любой моделью.

5.2.10 Генерализация денойзинга на разные SNR

Существующие аудиосигналы могут содержать различный по SNR шум, поэтому необходимо рассмотреть, как меняется качество распознавания при различных уровнях шума. Для экспериментов в данном случается рассматривается следующая постановка:

- В качестве датасета для обучения и валидации используется русскоязычный CommonVoice, зашумленный белым шумом с SNR в диапазоне от 5 до 25
- Обучается facebook denoiser на выходах русскоязычного wav2vec
- Для сравнения производится дообучение русскоязычного wav2vec
- Производится тестировани оригинальной модели wav2vec, wav2vec с денойзером, обученным методом DASR, и дообученного wav2vec при различных SNR

Результаты данных экспериментов представлены на рисунке 15.

Как видно из графиков, при сильной зашумленности (SNR=5) разница в качестве между оригинальной моделью wav2vec и моделью с денойзингом составляет

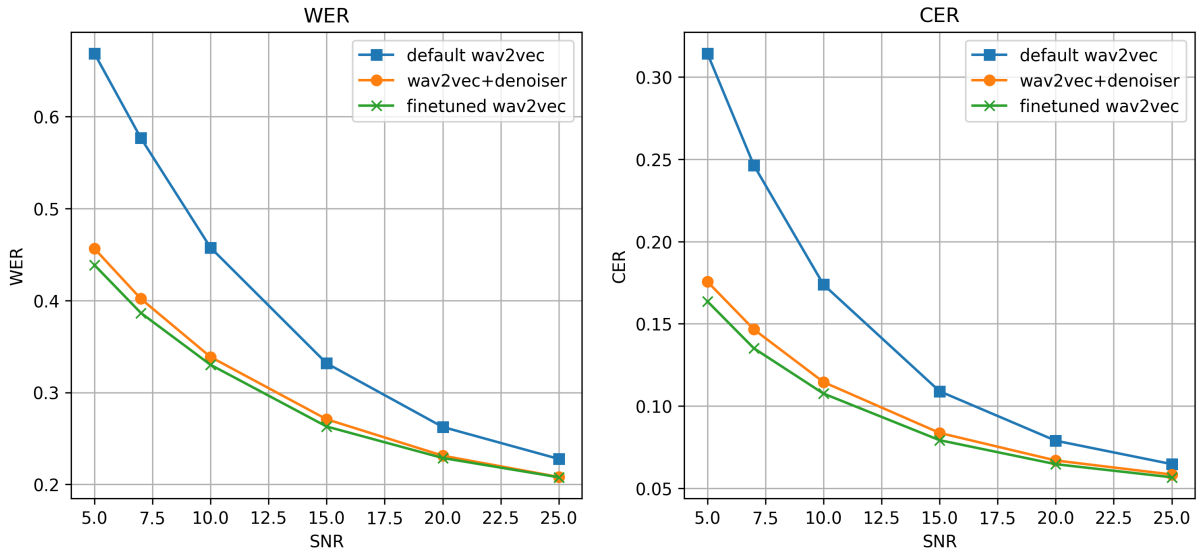

Рис. 15: Зависимость WER и CER от SNR

порядка 20 единиц WER и 15 единиц CER. По мере уменьшения уровня шума разница в качестве становится меньше, однако даже при практически отсутствующей зашумленности (SNR=20 и больше), качество модели с денойзингом оказывается лучше, чем у оригинальной, что говорит об эффективности предложенного метода даже в тех случаях, когда в речи не наблюдается шума.

Кроме того, следует отметить, что при всех уровнях SNR качество модели с денойзингом практически совпадает с качеством дообученной модели, что подтверждает эффективность денойзинга в методе DASR и возможность достигать с помощью него максимально достижимого качества распознавания.

5.2.11 Выводы

Анализ результатов проведенных экспериментальных исследований позволяет сделать следующий выводы:

- Предложена методика создания зашумленных датасетов с заданным уровнем SNR
- Дообучение моделей распознавания на зашумленных данных приводит к значительному улучшению метрик WER и CER
- Применение недообученных денойзеров ухудшает качество распознавания, так как они существенно изменяют аудиосигнал, отдаляя его от входного распределения модели распознавания

- Дообученные методом DASR денойзеры улучшают качество распознавания речи. Эксперименты показали, что дообучение денойзеров приводит к метрикам, сравнимым с метриками дообучения модели распознавания, а иногда даже превосходит их

- Обучение денойзеров с нуля показало худшие результаты по сравнению с использованием предобученных моделей. Это указывает на то, что предобученные денойзеры содержат релевантную информацию для улучшения качества распознавания

- Добавление функции потерь денойзинга к CTC-функции ухудшает метрики распознавания. Это подтверждает, что несмотря на схожесть задач, для распознавания речи не требуется строго следовать постановке задачи денойзинга

- Коррекция транскрибации методом DASR с использованием моделей ruT5 и YandexSpeller улучшает WER, но в случае с ruT5 ухудшает CER. YandexSpeller показал себя лучше в этом аспекте, так как улучшает обе метрики. Комбинирование денойзинга и коррекции транскрибации значительно снижает WER, что демонстрирует эффективность интеграции этих методов

- Проведенные тесты на различных датасетах показали, что денойзеры, обученные на CommonVoice методом DASR, сохраняют свою эффективность и на других датасетах, подтверждая свою способность к генерализации. Это свидетельствует о высокой универсальности предложенного подхода к денойзингу DASR и его применимости в различных речевых доменах

- Результаты экспериментов показывают, что денойзеры, обученные на одной модели распознавания методом DASR, могут эффективно применяться к другим моделям, демонстрируя лишь незначительное снижение качества относительно оригинального для модели денойзера. Это свидетельствует о высокой степени генерализации денойзеров, которые, изменяя входной аудиосигнал, улучшают его распознаваемость независимо от конкретной модели распознавания.

- Эксперименты показали, что денойзинг показывает сравнимую с дообучением эффективность как при сильной зашумленности, так и оказывается эффективнее, чем оригинальная модель распознавания даже при отсутствии шума

6 Заключение

В данной работе была рассмотрена задача распознавания речи в условиях сильной зашумленности. Эта задача была разделена на две подзадачи: детекция голосовой активности в реальном времени и распознавание речи в условиях сильной зашумленности.

Для детекции голосовой активности в реальном времени разработан алгоритм RealVADR, основанный на существующих нейросетевых моделях. Исследовано, как параметры предложенного алгоритма влияют на качество распознавания, и показано, что качество онлайн-распознавания с его помощью сравнимо с качеством офлайн-распознавания как на публичных датасетах, таких как CommonVoice, так и на специально собранном в шумной робототехнической среде датасете, предложенном в данной работе.

Для задачи распознавания речи предложен оригинальный метод DASR по адаптации нейросетевых моделей денойзинга. Суть подхода заключается в дообучении денойзеров с использованием функции потерь от модели распознавания без изменения самой модели распознавания. Для дообучения денойзеров предложена методика формирования обучающих датасетов, и показано, что качество распознавания, достигаемое с помощью денойзеров, практически совпадает с максимально достижимым качеством распознавания. Проведены эксперименты по генерализации денойзеров, которые показали, что денойзеры хорошо адаптируются как к новым датасетам, так и к другим моделям распознавания, для которых они изначально не предназначены. Это делает их универсальным инструментом для повышения качества распознавания в условиях сильной зашумленности. Также показано, что использование денойзеров улучшает качество распознавания не только в условиях сильной зашумленности, но и в условиях практически отсутствующего шума. Кроме того, рассмотрено использование моделей коррекции транскрибации, и показано, что их применение совместно с денойзерами значительно повышает качество распознавания, позволяя достичь результатов, превосходящих качество оригинальной модели распознавания почти в два раза.

Дальнейшим развитием предложенного метода DASR является его обобщение на другие датасеты и модели распознавания и денойзинга, а также его применение для улучшения качества распознавания в малоресурсных языках.

Список литературы

1. *Bourlard H.*, *Morgan N.* Connectionist Speech Recognition: A Hybrid Approach. — 01.1994. — ISBN 978-1-4613-6409-2. — DOI: 10.1007/978-1-4615-3210-1.
2. *Cieri C.*, *Miller D.*, *Walker K.* The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text // Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04) / под ред. M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva. — Lisbon, Portugal : European Language Resources Association (ELRA), 05.2004. — URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/767.pdf.
3. The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments / M. Lincoln, I. McCowan, J. Vepa, H. Maganti // IEEE Workshop on Automatic Speech Recognition and Understanding, 2005. — 2005. — С. 357—362. — DOI: 10.1109/ASRU.2005.1566470.
4. *Fujimoto M.*, *Watanabe S.*, *Nakatani T.* Voice activity detection using frame-wise model re-estimation method based on Gaussian pruning with weight normalization. // INTERSPEECH. — 2010. — С. 3102—3105.
5. Noise robust voice activity detection using features extracted from the time-domain autocorrelation function / H. Ghaemmaghami, B. Baker, R. Vogt, S. Sridharan // Proceedings of the 11th Annual Conference of the International Speech Communication Association. — International Speech Communication Association. 2010. — С. 3118—3121.
6. Using Spectral Fluctuation of Speech in multi-feature HMM-based voice activity detection / M. Espi, S. Miyabe, T. Nishimoto, N. Ono, S. Sagayama // Twelfth Annual Conference of the International Speech Communication Association. — 2011.
7. Speech enhancement based on deep denoising autoencoder / X. Lu, Y. Tsao, S. Matsuda, C. Hori // Proc. Interspeech 2013. — 2013. — С. 436—440. — DOI: 10.21437/Interspeech.2013-130.
8. The IBM speech activity detection system for the DARPA RATS program. / G. Saon, S. Thomas, H. Soltau, S. Ganapathy, B. Kingsbury // Interspeech. — 2013. — С. 3497—3501.

9. Deep Speech: Scaling up end-to-end speech recognition / A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, A. Y. Ng. — 2014. — arXiv: 1412.5567 [cs.CL].

10. A regression approach to speech enhancement based on deep neural networks / Y. Xu, J. Du, L.-R. Dai, C.-H. Lee // IEEE/ACM Trans. Audio, Speech and Lang. Proc. — 2015. — Янв. — Т. 23, № 1. — С. 7—19. — ISSN 2329-9290. — DOI: 10.1109/TASLP.2014.2364452. — URL: https://doi.org/10.1109/TASLP.2014.2364452.

11. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin / D. Amodei [и др.]. — 2015. — arXiv: 1512.02595 [cs.CL].

12. Librispeech: An ASR corpus based on public domain audio books / V. Panayotov, G. Chen, D. Povey, S. Khudanpur // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — 2015. — С. 5206—5210. — DOI: 10.1109/ICASSP.2015.7178964.

13. Feature Learning with Raw-Waveform CLDNNs for Voice Activity Detection. / R. Zazo, T. N. Sainath, G. Simko, C. Parada // Interspeech. — 2016. — С. 3668—3672.

14. *Montacié C.*, *Caraty M.-J.* Prosodic Cues and Answer Type Detection for the Deception Sub-Challenge. // INTERSPEECH. — 2016. — С. 2016—2020.

15. *Bai L.*, *Zhang Z.*, *Hu J.* Voice activity detection based on deep neural networks and Viterbi // IOP Conference Series: Materials Science and Engineering. Т. 231. — IOP Publishing. 2017. — С. 012042.

16. *Pascual S.*, *Bonafonte A.*, *Serrà J.* SEGAN: Speech Enhancement Generative Adversarial Network // CoRR. — 2017. — Т. abs/1703.09452. — arXiv: 1703.09452. — URL: http://arxiv.org/abs/1703.09452.

17. *Rethage D.*, *Pons J.*, *Serra X.* A Wavenet for Speech Denoising // CoRR. — 2017. — Т. abs/1706.07162. — arXiv: 1706.07162. — URL: http://arxiv.org/abs/1706.07162.

18. *Jassim W. A.*, *Harte N.* Voice activity detection using neurograms // 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). — IEEE. 2018. — С. 5524—5528.

19. *Shao Y.*, *Lin Q.* Use of pitch continuity for robust speech activity detection // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE. 2018. — C. 5534—5538.

20. A Comparative Study on Transformer vs RNN in Speech Applications / S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang // CoRR. — 2019. — T. abs/1909.06317. — arXiv: 1909.06317. — URL: http://arxiv.org/abs/1909.06317.

21. *Baevski A.*, *Schneider S.*, *Auli M.* vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations // CoRR. — 2019. — T. abs/1910.05453. — arXiv: 1910.05453. — URL: http://arxiv.org/abs/1910.05453.

22. Common voice: A massively-multilingual speech corpus / R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber // arXiv preprint arXiv:1912.06670. — 2019.

23. Improving RNN Transducer Modeling for End-to-End Speech Recognition / J. Li, R. Zhao, H. Hu, Y. Gong // 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). — 2019. — C. 114—121. — DOI: 10.1109/ASRU46091.2019.9003906.

24. Music source separation in the waveform domain / A. Défossez, N. Usunier, L. Bottou, F. Bach // arXiv preprint arXiv:1911.13254. — 2019.

25. wav2vec: Unsupervised Pre-training for Speech Recognition / S. Schneider, A. Baevski, R. Collobert, M. Auli // CoRR. — 2019. — T. abs/1904.05862. — arXiv: 1904.05862. — URL: http://arxiv.org/abs/1904.05862.

26. *Defossez A.*, *Synnaeve G.*, *Adi Y.* Real time speech enhancement in the waveform domain // arXiv preprint arXiv:2006.12847. — 2020.

27. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer / C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu // Journal of Machine Learning Research. — 2020. — T. 21, № 140. — C. 1—67. — URL: http://jmlr.org/papers/v21/20-074.html.

28. Unsupervised Cross-lingual Representation Learning for Speech Recognition / A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli // CoRR. — 2020. — T. abs/2006.13979. — arXiv: 2006.13979. — URL: https://arxiv.org/abs/2006.13979.

29. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations / A. Baevski, H. Zhou, A. Mohamed, M. Auli // CoRR. — 2020. — T. abs/2006.11477. — arXiv: 2006.11477. — URL: https://arxiv.org/abs/2006.11477.

30. *Karpov N.*, *Denisenko A.*, *Minkin F.* Golos: Russian Dataset for Speech Research. — 2021. — URL: https://arxiv.org/abs/2106.10161.

31. *Tan K.*, *Zhang X.*, *Wang D.* Deep Learning Based Real-time Speech Enhancement for Dual-microphone Mobile Phones // IEEE/ACM Trans Audio Speech Lang Process. — United States, 2021. — Май. — Т. 29. — С. 1853—1863.

32. Unsupervised Speech Recognition / A. Baevski, W. Hsu, A. Conneau, M. Auli // CoRR. — 2021. — T. abs/2105.11084. — arXiv: 2105.11084. — URL: https://arxiv.org/abs/2105.11084.

33. *Zubarev E.*, *Moskalets T.*, *SOVA.ai.* SOVA RuDevices Dataset: free public STT/ASR dataset with manually annotated live speech. — 2021. — https://github.com/sovaai/sova-dataset.

34. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech / A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, A. Bapna // arXiv preprint arXiv:2205.12446. — 2022. — URL: https://arxiv.org/abs/2205.12446.

35. Robust Speech Recognition via Large-Scale Weak Supervision / A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever. — 2022. — arXiv: 2212.04356 [eess.AS].

36. A family of pretrained transformer language models for Russian / D. Zmitrovich, A. Abramov, A. Kalmykov, M. Tikhonova, E. Taktasheva, D. Astafurov, M. Baushenko, A. Snegirev, T. Shavrina, S. Markov [и др.] // arXiv preprint arXiv:2309.10931. — 2023.

37. Llama 2: Open Foundation and Fine-Tuned Chat Models / H. Touvron [и др.]. — 2023. — arXiv: 2307.09288 [cs.CL].

38. Mistral 7B / A. Q. Jiang [и др.]. — 2023. — arXiv: 2310.06825 [cs.CL].

39. *Петряшин И.*, *Юдин Д.* Нейросетевая детекция голосовой активности для распознавания речи в реальном времени // Труды МФТИ 2023. —.